DOCUMENT RESUME

ED 301 571 TM 011 274

AUTHOR	Bennett, Randy Elliot; And Others
TITLE	The Psychometric Characteristics of the GRE General Test for Three Handicapped Groups. Studies of Admissions Testing and Handicapped People.
INSTITUTION	Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
SPONS AGENCY	College Entrance Examination Board, New York, N.Y.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-86-6
PUB DATE	Feb 86
NOTE	85p.
PUB TYPE	Reports - Research/Technical (143) -- Collected Works - Serials (022)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Comparative Analysis; Graduate Study; Higher Education; Item Analysis; *Physical Disabilities; *Psychometrics; Test Bias; *Test Reliability; *Timed Tests; *Visual Impairments
IDENTIFIERS	Differential Item Performance; *Graduate Record Examinations

ABSTRACT

The psychometric characteristics of the Graduate Record Examinations General Test (GRE-GT) were studied for three handicapped groups. Experimental subjects took the GRE-GT between October 1981 and June 1984; they include: (1) 151 visually-impaired students taking large-type, extended-time administrations; (2) 188 visually-impaired students taking timed, regular-type administrations; (3) 108 physically-handicapped students taking timed, regular-type administrations; and (4) 441,654 non-handicapped students. Data on a non-handicapped subset (N=20,499) who took the GRE-GT in October 1981 and April 1982 were included. Psychometric characteristics studied included the level of test performance, test completion, test reliability, and unexpected differential item performance. Results indicate that the performance of visually-handicapped groups closely approximates that of non-disabled students. However, physically-handicapped students performed substantially lower than did the comparison group of non-handicapped students on two of the GRE-GT scales. Data on test completion showed evidence of possible disadvantage to the two groups of disabled examinees taking timed administrations. Test reliability analysis via internal consistency and parallel-forms methods indicated no practical differences in measurement precision across groups. In terms of both reliability and item behavior, the psychometric characteristics of the GRE-GT appear comparable for handicapped and non-handicapped groups. Nineteen data tables and 11 sample GRE-GT questions are provided. An annotated bibliography of five reports of the "Studies of Admissions Testing and Handicapped People" is appended. (TJH)

9033

RR-86-6

ED301571

THE PSYCHOMETRIC CHARACTERISTICS OF THE GRE GENERAL TEST FOR THREE HANDICAPPED GROUPS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

HELEN C. WEIDENHILLER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Randy Elliot Bennett
Donald A. Rock
and
Thomas Jirele

February 1986

BEST COPY AVAILABLE

Report No. 6
Studies of Admissions Testing and Handicapped People
A Project Sponsored by

College Entrance Examination Board
Educational Testing Service
Graduate Record Examinations Board

TM 011 274

Studies of Admissions Testing and Handicapped People

Most admissions testing programs have long made accommodations for handicapped examinees, though practices have varied across programs and limited research has been undertaken to evaluate such test modifications. Regulations under Section 504 of the Rehabilitation Act of 1973 impose new requirements on institutional users, and indirectly on admissions test sponsors and developers, in order to protect the rights of handicapped persons. The Regulations have not been strictly enforced since many have argued that they conflict with present technical capabilities of test developers. In 1982, a Panel appointed by the National Research Council released a detailed report and recommendations calling for research on the validity and comparability of scores for handicapped persons.

Due to a shared concern for these issues, College Board, Educational Testing Service, and Graduate Record Examinations Board initiated a series of studies in June 1983. The primary objectives are:

To develop an improved base of information concerning the testing of handicapped populations.

To evaluate and improve wherever possible the accuracy of assessment for handicapped persons, especially test scaling and predictive validity.

To evaluate and enhance wherever possible the fairness and comparability of tests for handicapped and nonhandicapped examinees.

This is one of a series of reports on the project, which will continue through 1986. Opinions expressed are those of the authors. See Appendix for an annotated bibliography of earlier reports of the series.

ETS Research Report RR-86-6

The Psychometric Characteristics of the GRE General Test for Three Handicapped Groups

Randy Elliot Bennett

Donald A. Rock

and

Thomas Jirele

Educational Testing Service

Princeton, NJ 08541

February 1986

Copyright (c) 1986 by Educational Testing Service. All rights reserved.

The ETS logo, Educational Testing Service, GRE, and Graduate Record Examinations Board, are registered trademarks of Educational Testing Service.

College Board and the acorn logo are registered trademarks of the College Entrance Examination Board.

Abstract

This study examined the psychometric characteristics of the Graduate Record Examinations (GRE) General Test for three handicapped groups. Information about test characteristics is central to judging the accuracy and fairness of General Test scores for handicapped students.

Four psychometric characteristics were studied: level of test performance, test completion, test reliability, and unexpected differential item performance. Psychometric comparisons were made between nonhandicapped samples and each of three handicapped groups: visually impaired students taking large-type, extended-time administrations; visually impaired students taking timed, regular-type administrations; and physically handicapped students taking timed, regular-type administrations.

Results of the study showed the performance of the visually handicapped groups to closely approximate that of nondisabled students. Physically handicapped students, however, performed substantially lower than the comparison group on two of the three General Test scales.

Data on test completion showed evidence of possible disadvantage to the two groups of disabled examinees taking timed administrations. Evidence that handicapped students were somewhat less likely to complete the test was found for both groups on the Analytical sections and for one group on the first Quantitative section. Though these scores may, as a result, underestimate developed scholastic ability, any

conclusion on this account must await the findings of further research.

Test reliability was analyzed using both internal consistency and parallel-forms methods. Neither method revealed any practical differences in measurement precision across groups.

Because of the large number and general unreliability of individual item-performance contrasts, unexpected differential item performance was examined through a two-stage procedure. The first stage centered on the performance of item clusters. Items composing clusters showing questionable performance were then individually examined. This two-stage procedure revealed only two instances--localized to visually impaired students taking the large-type test--of differential item performance.

In terms of both reliability and item behavior, the psychometric characteristics of the GRE General Test appear comparable for the handicapped and nondisabled groups studied. Though somewhat less comparable, test completion and performance levels also were similar across groups. Further studies of General Test scores--in particular, their factor structure and predictive validity--should provide additional information about their meaning for handicapped students.

Acknowledgements

Appreciation is expressed to Joan Chase, Neil Dorans, June Mullins, Catherine Nelson, Donald Powers, Lawrence Stricker, and Warren Willingham for their reviews of an earlier version of this manuscript, and to Susan Vitella for providing general information about the GRE and its services to handicapped students. Gratitude is also expressed to the members of the Graduate Record Examinations Board Research Committee and the College Board Joint Staff Research and Development Committee.

Table of Contents

The Graduate Record Examinations.....1

Subjects.....9

Results.....13

Level of Performance.....14

Test Completion.....19

Reliability.....24

Unexpected Differential Performance.....27

Item clusters.....30

Individual items.....36

Summary.....43

References.....46

Footnote.....49

Tables.....50

Figures.....69

Appendix - Annotated Bibliography of Previous Reports of the "Studies of Admissions Testing and Handicapped People" 75

In 1983, the College Board, Educational Testing Service (ETS), and the Graduate Record Examinations Board initiated a joint project, "Studies of Admissions Testing and Handicapped People," in response to a call by a National Academy of Sciences Panel for further research into the use of college and graduate admissions tests for handicapped people (Sherman & Robinson, 1982). As part of that joint research effort, this study presents information on the psychometric characteristics of the Graduate Record Examinations (GRE) General Test for three groups of handicapped examinees. The study reports data on the level of performance, test completion reliability, and unexpected differential item performance for these groups. These data, in particular those on reliability and differential performance, are fundamental to evaluating the extent to which the General Test fairly and accurately measures the developed scholastic abilities of handicapped students.

The Graduate Record Examinations

The Graduate Record Examinations are developed and administered by ETS under policies determined by the Graduate Record Examinations Board, an independent board affiliated with the Association of Graduate Schools and the Council of Graduate Schools in the United States (Educational Testing Service, 1984). The Examinations are used by graduate and professional schools, and by fellowship panels, to supplement undergraduate records and other indicators of a candidate's readiness for graduate study.

The Graduate Record Examinations include a General Test and Subject Tests. The General Test is a multiple-choice examination designed to measure broad, developed abilities generally required for success in graduate work. In contrast, the Subject Tests are meant to measure knowledge and understanding of content basic to study in specific fields. The Subject Tests are offered in 17 fields including biology, computer science, economics, education, engineering, French, history, music, psychology, and sociology.

The General Test is made up of three sections designated Verbal, Quantitative, and Analytical. The Verbal section is intended to test the examinee's ability to reason with words in solving problems. The section is composed of 76 items falling into four categories: analogies (18 questions), antonyms (22 questions), sentence completion (14 questions), and reading comprehension (22 questions).

Analogies items are meant to assess the ability to recognize relationships among words and the concepts they represent, and to recognize when these relationships are parallel (Educational Testing Service, 1984). The kinds of relationships found in such items may include size, kind, continuity, and degree. The purpose of antonym questions is to measure both the strength of one's vocabulary and the ability to reason from a given concept to its opposite. The words used in these questions are generally confined to nouns, verbs and adjectives. Sentence Completion questions

are designed to assess the ability to recognize words or phrases that logically and stylistically complete the meaning of a sentence. These items provide a context within which to analyze the function of words as they relate to and combine with one another to form a meaningful unit of discourse.

The last Verbal item type, reading comprehension, explores the candidate's ability to recognize explicitly-stated elements and assumptions underlying statements or arguments in the passage as well as their implications (Educational Testing Service, 1984). Four passages representing different subject matter areas are presented on each General Test. The subject matter areas from which passages are drawn include the humanities, social sciences, biological sciences, and the physical sciences. Six types of questions are posed about these passages. These questions inquire about: (1) the main idea or primary purpose of the passage, (2) information explicitly stated in it, (3) information or ideas implied or suggested by the author, (4) possible application of the author's ideas to other situations, (5) the author's logic, reasoning or persuasive techniques, and (6) the author's tone or attitude as revealed by the language used.

Examples of each GRE-Verbal item type are presented in Figure 1.

Insert Figure 1 about here

The Quantitative section of the General Test is meant to measure basic mathematical skills, understanding of elementary mathematical concepts, and ability to reason quantitatively (Educational Testing Service, 1984). The content of quantitative questions is balanced among arithmetic, algebra, and geometry. Sixty questions are presented divided into three formats: quantitative comparison (30 questions), discrete quantitative (20 questions), and data interpretation (10 questions).

Quantitative comparison items typically present two quantities. The test candidate must examine the quantities and select from four options the one that best describes the relationship between the two amounts. Quantitative comparisons test the ability to reason quickly and accurately about the relative sizes of two quantities or to perceive that not enough information is provided to make a decision (Educational Testing Service, 1984).

Discrete quantitative items contain all the information needed for answering the question except for basic mathematical knowledge. Many such items require little more than simple manipulation of quantities. Others are presented in a text format and require the examinee to read, understand, and solve a word problem. All discrete

quantitative items are presented in the standard multiple-choice format.

The third Quantitative item type, data interpretation, typically appears in sets. Each set of items is based on data presented in tables or graphs. Data interpretation questions are meant to test one's ability to synthesize information, to select appropriate data for answering a question, or to determine that sufficient information is not provided (Educational Testing Service, 1984).

Examples of the quantitative comparison, discrete quantitative, and data interpretation item types are presented in Figure 2.

Insert Figure 2 about here

The third component of the General Test is the Analytical section. This section was added in 1977, restructured in October 1981, and administered on an experimental basis through Fall 1985, at which time it became a permanent part of the General Test. The section is composed of two item types: analytical reasoning (38 items) and logical reasoning (12 items). Analytical reasoning questions are designed to measure the ability to analyze a given structure of arbitrary relationships and to deduce new information from that structure (Educational Testing Service, 1984). These items are presented in groups of three or more with each group based on a different set of

conditions describing a fictional situation. Logical reasoning items focus on the ability to understand and to analyze relationships among arguments or parts of an argument. These items are based on a short prose passage, usually a separate passage for each item.

Examples of the two Analytical item types are presented in Figure 3.

Insert Figure 3 about here

When administered, the GRE General Test is divided into seven separately timed, 30-minute sections: two Verbal, two Quantitative, two Analytical, and one experimental that does not count toward the student's score. The two sections of each kind are constructed according to similar specifications: the same number, difficulty, and types of items appear. Within a section, items of a similar format are grouped together.

National administrations of the GRE are offered five times a year. Differences in mean scores across these administrations vary substantially, with candidates taking the test early in the academic year typically scoring higher than those at later administrations. Those taking the examination earlier include applicants for National Science Foundation and other fellowship programs, and students applying for admission to competitive master's or doctoral programs. Those taking the test later more often apply to

less competitive "rolling" admissions programs that accept students at several points during the academic year.

Special administrations for handicapped students are available to those with visual, physical, hearing, and learning disabilities (Educational Testing Service, 1984). Alternative test formats as well as modified conditions for administering the test are offered. Alternative formats include braille (1972 Nemeth code for the Quantitative section), large print, and cassette. The cassette edition is accompanied by diagrams, charts, and drawings in either large-type or raised-line format, depending upon the preference of the examinee; the illustrations also are described on tape by the reader. A large print answer sheet is available with all alternative formats. Modified conditions consist of the use of a reader, amanuensis (recorder), separate testing room, or extra time (up to six hours). The use of adaptive devices, such as magnifying glasses, overhead readers, Optacons, and other aids also is permitted.

Results of General Test administrations are reported for Verbal, Quantitative, and Analytical performance, each on a 200 to 800 standard-score scale with a mean of 500 and standard deviation of 100. Individual scores are always reported with a zero in the one's column so that in actuality the scale has 61 as opposed to 601 score units. The scale is based on the performance of examinees taking the test in 1952 (Conrad, Trismen, & Miller, 1977); the

performance of all subsequent groups is statistically equated to that original administration. Hence, the means and standard deviations of groups taking the test have deviated over the years from their original values, but the meaning of scores has stayed the same. The effect on scores of nonstandard administrations is not known. Therefore, all such scores are accompanied by a statement emphasizing the importance of considering other indicators of academic achievement in the admissions process.

Much is known about the psychometric characteristics of the General Test in the larger population of students attending graduate and professional schools. Little data on the characteristics of the test for handicapped students, however, exist (Bennett, Ragosta, & Stricker, 1984). For the nonhandicapped population, correlation coefficients with first-year grades averaged across 229 graduate departments were reported to be .23 for the Verbal score, .22 for the Quantitative score and .26 for the total test (Burton & Turner, 1983). (No predictive data have been published yet for the current version of the Analytical scale.) The median predictive validity for undergraduate grade-point average is .27, about equivalent to that for the total test. When both predictors are combined, the validity coefficient increases to .34.

As with most averages, these median coefficients mask variation. The predictive validity of the GRE varies as a function of discipline, institutional characteristics

(selection rules, grading standards, educational program), academic year, student population, and other factors. For example, the median correlation of Quantitative score with first-year grades in humanities departments was .13 (32 departments), while the median correlation with grades in mathematical and physical sciences was .28 (48 departments).

Subjects

During the period from October 1981 through June 1984, 663,319 examinees took the General Test in domestic administrations and 14,142 of these (about 2%) identified themselves as handicapped. Of this latter number, 13,284 sat for the standard, national exam and 858 took special administrations. Those taking special administrations were divided among alternate formats, taken by 414 students, and the regular type edition of the test with extended time, administered to 444 individuals. (1)

During this period, the Graduate Record Examinations program administered one form of the General Test, designated as 3DGR3, in alternate formats. This form was also used, in addition to eleven other regular-type forms, for handicapped students requiring only extra time. In addition, the form was administered (along with 18 others) under standardized conditions to students participating in national administrations.

Because of its use in both special and standard administrations, 3DGR3 was selected for analysis in the current study. In addition to its use in special formats

from October 1981 to June 1984, the regular-type edition of 3DGR3 was administered with extended time in October 1981. A modification of the form differing only in the revision of two items, K-3DGR3, was given in the same manner in June 1982. The same forms were also administered nationally under standardized conditions in October 1981 and April 1982, respectively.

Table 1 indicates the numbers of handicapped students who took these two forms under various conditions. In the table, entries are collapsed across 3DGR3 and K-3DGR3. The combined form is referred to as C-3DGR3. The primary virtue of the combined form is that it increases the sizes of the various handicapped groups. This increase both permits more groups to be studied and raises the level of confidence that can be placed in the stability of results derived from a group.

Insert Table 1 about here

Students for whom proficiency in English was questionable are excluded from the table. These students were excluded to avoid confounding differences in the operation of the test owing to handicap with those related to language proficiency. The excluded group of students consisted of (1) those taking foreign administrations, (2) noncitizens and those with unknown citizenship who indicated on the GRE registration form that they communicated better

in a language other than English or who failed to respond to the question, and (3) citizens indicating that English was not their best language of communication. Deaf students, many of whom communicate best in sign language, were not subjected to the language exclusion criterion.

As the table indicates, handicapped students taking C-3DGR3 fell into six major groups: visually impaired, hearing impaired, physically handicapped (i.e., those indicating paraplegia, other neurological or orthopedic impairments, or a permanent physical disability), learning disabled, multiply handicapped, and a group labeled other, composed of students with unspecified handicaps. The definitions and characteristics of visual, hearing, physical and learning disabilities are discussed in an earlier report (Bennett & Ragosta, 1984). For purposes of the current study, students were considered to belong to one of these disability groups on the basis of self-report made as part of the General Test registration process. While inaccurate self-reports of handicapped group membership would be expected to affect the meaning of research results, there is no indication that such inaccuracies are present in any number.

The major administrative formats offered for C-3DGR3 included alternate formats (i.e., braille, large type, cassette, and multiple, or combined, editions), regular type with extra time, and the national administration in which the test is taken with nonhandicapped students under

standard (timed) conditions. All alternate format administrations included extended time.

The largest number of administrations (454) was taken by visually impaired students and the most frequently used format was the standard, national administration (537). Students with unspecified impairments represented the second largest disability group (130) and large-type the second most frequently-used format (185). Of the 36 possible test format-by-disability group combinations, the two largest were visually impaired students taking the national administration (188) and visually impaired students taking large-type administrations (151).

In addition to these two groups, two other format-by-group combinations have numbers of students (100 or more) sufficient to support dependable results and justify further study. These groups are the physically handicapped-national administration and other-national administration students. Of these groups, the other-national administration sample was excluded because its composition was unknown.

Table 2 lists the acronyms used to denote the three disability groups selected for study. Also listed are the sample sizes for each group.

Insert Table 2 about here

To evaluate properly the psychometric characteristics of the GRE General Test for these three disability groups

some reference, or "standard," population is needed. Without such a population, the typical behavior of the test cannot be known and any departures from this behavior by subpopulations cannot be detected. In the present study, two standard groups are used. The first group is composed of 441,654 students taking administrations of the General Test during the October 1981 to June 1984 time period. The second standard group is a subset of this larger group and consists of 20,499 individuals taking C-3DGR3 under typical testing conditions in October 1981 and April 1982. As with the handicapped samples, students with questionable levels of English language competency have been removed from these groups.

Table 3 lists the mean Verbal, Quantitative, and Analytical scores for examinees taking C-3DGR3 and those taking all forms of the General Test (including C-3DGR3) during the October 1981 to June 1984 period. These groups are denoted by the acronyms NHF and NHA, respectively. As the table shows, only minor differences exist between the two samples.

Insert Table 3 about here

Results

Results are reported for level of performance, test completion, reliability, and unexpected differential item behavior. Because General Test scores are reported on the

same scale regardless of the form taken, the first analysis is based upon scores from the complete versions of 3DGR3 and K-3DGR3. For purposes of the other analyses, the two items that distinguish 3DGR3 from K-3DGR3 are omitted, yielding a single common set of items taken by all subjects.

Level of Performance

Table 4 lists scaled score means and standard deviations for the performance of the three handicapped groups on the Verbal, Quantitative, and Analytical sections of C-3DGR3. To facilitate comparison with students typically taking the General Test, the standardized difference between each handicapped group's mean score and the appropriate reference group is presented. (This difference is calculated by subtracting the scaled score mean of the handicapped group from the comparable value for the reference group and then dividing by the reference group's standard deviation.) Because the group of visually impaired-large type (VIL) examinees represents an accumulation of administrations taken over the October 1981 to June 1984 period, the performance of these individuals is compared to nondisabled students taking the General Test during this time. The remaining two handicapped groups are derived from national administrations offered in October 1981 and April 1982. These groups are compared with nondisabled examinees taking contemporaneous administrations.

Insert Table 4 about here

Review of Table 4 suggests performance substantially similar to the nonhandicapped reference groups for two of the three disability samples: visually impaired-large type (VIL) students achieve mean scaled scores from .11 standard deviations below to .28 standard deviations above the mean, while visually impaired-national administration (VIN) examinees show values between -.13 and -.10 standard deviation units. Physically handicapped examinees (PHN) scored lower, from .43 to .03 standard deviations below the reference group. These ranges are roughly consistent with those reported for the performance of physically handicapped and visually impaired examinees on undergraduate admissions tests (Bennett, Ragosta, & Stricker, 1984; Bennett, Rock, & Kaplan, 1985).

The data in Table 4 can also be interpreted with respect to the three General Test sections. On Verbal, the handicapped samples closely approximate reference group performance. Mean scores for this scale range from -.03 (for PHN) to .12 (for VIL) standard deviations from the mean. For the Quantitative section, only the physically handicapped-national administration group (PHN) evidences performance substantially below (by .43 standard deviations) the reference average. Finally, scores for Analytical,

while bracketing the reference group mean, are somewhat dispersed, ranging from -.31 (for PHN) to .28 (for VIL).

In addition to differences in mean performance, the degree of variability evidenced is, in one case, noteworthy. On the Analytical section, the visually impaired-large type (VIL) group shows an unusually large standard deviation with respect to the reference group (i.e., 154 vs. 125). No explanation for this greater variation is evident.

Several factors may account for the few appreciable differences in mean test performance observed between handicapped and nondisabled groups. Aside from true differences in developed abilities or pseudo-differences caused by inadequacies in the test, observed disparities in scores may be due to in part to differences in the distribution of such background characteristics as sex or undergraduate major. For example, in national samples taking the test during the October 1981 to June 1984 period, the sexes achieved generally similar mean scores on Verbal and on the Analytical scale. On Quantitative, however, females typically attained scores about .6 standard deviation units lower than males (Goodison, 1983; Smith, 1984, 1985). Assuming that visual and physical handicaps are not linked intrinsically to sex, it is possible for a disability group to evidence mean performance different from the reference population not because of disability, but because its sex composition differs substantially from the base group. (For the same reason, it is possible for true

differences in performance between groups to be masked, as when a handicapped group has a greater proportion of males than the reference group.)

Table 5 shows the sex distribution of each handicapped group in comparison with the appropriate reference population. As the table indicates, the reference populations have male-to-female ratios of approximately 4:6. Two of the three handicapped samples diverge somewhat from this composition: the physically handicapped-national administration group (PHN), which is slightly more male than female, and the visually impaired large-type group (VIL), which has a greater proportion of females than the reference population. Because of the small sample sizes in the handicapped groups, these differences may not be dependable. If real, however, they suggest that the already-low mean Quantitative score of the physically handicapped group (PHN) would be somewhat lower if the group had the same sex distribution as the reference group. On the other hand, the mean score of the visually impaired group (VIL), which already closely approximates the reference group value, might be higher given comparable sex distributions.

Insert Table 5 about here

As noted, undergraduate major is also commonly associated with General Test Performance. For example, for a national sample of examinees taking the General Test

during the 1981 to 1983 academic years, social science majors scored lowest of five undergraduate major groups on each of the General Test scales. Those with the highest mean Quantitative and Analytical scores--examinees reporting majors in the physical sciences--scored about 1.6 standard deviations higher on Quantitative and .5 higher on Analytical than social science majors. Those with the highest mean Verbal scores--humanities majors--performed .4 standard deviations above those specializing in the social sciences.

The relationship between the distribution of undergraduate major and differences in the test scores of handicapped and nondisabled groups is more conceptually complex than the corresponding relationship between sex and handicap. While sex group membership is inherent, handicapped students (like nondisabled ones) may choose a particular undergraduate major <u>because</u> they are more adept at it. Just the same, they may avoid a field because their disability makes it difficult for them to succeed. In such cases, the observed difference in test scores between handicapped and nondisabled groups may reflect both differences associated with the distribution of undergraduate major and real differences in developed abilities associated with handicap.

Table 6 shows the distribution for five categories of undergraduate major across handicapped and nondisabled groups. As the table suggests, the distribution of the

visually impaired-national administration (VIN) sample is very similar to the base group, the physically handicapped (PHN) sample diverges somewhat, and the visually impaired-large type (VIL) sample is most discrepant of all. Of particular interest are differences in the distribution of social and physical science majors: these majors are most related to performance on the Quantitative and Analytical scales, on which handicapped groups do least well. As the table indicates, social science majors are slightly overrepresented in the physically handicapped (PHN) group and more substantially overrepresented in the visually impaired-large type (VIL) sample; physical science majors appear slightly underrepresented in both visually impaired groups. Again, how disabilities affect the choice of these majors, and how the two characteristics in combination relate to total score differences, is unclear.

Insert Table 6 about here

Test Completion

Test completion statistics are typically used as a means of assessing the degree to which a test is speeded. In such uses, high rates of completion suggest that speed of work plays little role in test performance. Low rates imply the converse.

Unlike admissions tests such as the SAT, the GRE General Test is scored using the "rights only" method; no

penalty for guessing is imposed. It is, therefore, to the advantage of the examinee to complete every test item, and examinees are given this direction when the test is introduced. Most examinees should, knowing that time is about to elapse, randomly mark responses to any questions they have yet to answer. Given this fact, the relationship between test completion and speededness becomes clouded (Bejar, 1985). Are differences in completion rates between groups due to differences in the pace of taking the test or, alternatively, do the groups work at the same rate but comprehend the scoring instructions differently? Is one group more likely to randomly mark remaining questions while the other leaves those questions blank?

Under rights-only scoring, test completion rates cannot be related directly to speededness, but they may still provide information of value in evaluating the operation of a test. If rates across groups differ, the test may not be fairly representing the abilities of one group relative to another. For example, two groups of equal ability working at the same pace will receive different scores if they use differing end-of-test response strategies. Similarly, groups of equal ability will score differently if motor or visual impairments cause one group to work more slowly, thereby answering fewer questions in the same time.

Although differences in completion rates may indicate the presence of irrelevant variance in test scores, the absence of such differences does not necessarily guarantee

that scores are free of these influences. For instance, two groups of equal ability may work at different rates, the slower group impeded by motor impairments. Upon learning that time is about to elapse, both groups may complete the test by randomly marking responses to the remaining items. However, the faster group will likely receive the higher score as its members have had a chance to thoughtfully respond to more items than the slower-working group.

To assess the extent to which the test is completed by disabled students, two indices, the percent of students completing the section and the percent finishing 75% of the section, were computed and compared to those for nonhandicapped individuals taking C-3DGR3 in national administrations. These indices are jointly considered in the evaluation of test completion because in isolation neither is a fully satisfactory indicator. For example, the percent completing the section is based entirely on those responding to the last item. As such, the index cannot distinguish between those not reaching the item and those omitting the item for some reason. If the last item is omitted by one group to a greater degree than by another, the index may make differences in group test completion look more substantial than they really are.

Table 7 presents the ratio of each disability group index to its reference-group counterpart. Values of 1.00 indicate equal percentages completing the section or part section for both groups, while those above 1.00 suggest

greater completion rates for disabled students. When both indices are considered, several inferences are supported. First, almost without exception, the handicapped group participating in the special administrations program (i.e., the visually impaired-large type [VIL] sample) appears to complete the section or part-section in equal or slightly greater proportions than the reference group. Given that extended time is allowed this group and that the group scores above the reference group on two of the three General Test scales, it is likely that the performance of the group is not being hampered by lack of time or differences in test-taking strategy.

Insert Table 7 about here

Evidence of completion differences, however, is apparent for the two national administration groups. Slightly smaller proportions of the physically handicapped (PHN) sample reach the completion points (i.e., 75% of the test and the end of the test) than their nonhandicapped peers for the first Quantitative section and for the two Analytical sections. Slight differences on the Analytical sections likewise appear for the visually handicapped (VIN) group.

Are these disparities due to different end-of-test taking strategies? Table 8 presents the absolute proportions of students completing each General Test

section. For nondisabled students, these proportions range from 83 to 93 percent. Certainly if this group had conventionally used the expected random, end-of-test response strategy completion percentages would more closely approximate 100%.

Insert Table 8 about here

Are the observed differences due to speededness factors unrelated to scholastic ability? Certainly, such an interpretation is plausible. Because of their motor impairments, physically handicapped examinees probably take longer to record their answers, to perform the written computations necessary for some Quantitative problems, and to draw diagrams helpful in solving certain Analytical items. Similarly, visually impaired students may take longer to read test directions and items, and to locate the correct boxes on the answer sheet.

Finally, is it possible that differences in developed ability are responsible for observed disparities in test completion? The sections in which completion differences appear belong to the scales on which both physically and visually impaired examinees score lowest. Do these groups move more slowly through these sections simply because they are less adept in the abilities being measured?

It should be obvious that definitive explanations for these completion differences cannot be offered. Construct

and predictive validity studies should further explicate the extent to which handicapped students' General Test scores are differentially affected by such irrelevant sources of variance as speed of work and test taking strategy.

Reliability

Reliability refers to the precision or accuracy with which a test measures. Differences in the precision of measurement across groups can negatively impact upon the less accurately-measured group. For example, consider what might happen if a graduate admissions test measured less precisely for visually impaired than for sighted students. In this situation, the dispersion of the observed scores of visually impaired students around their true scores (i.e., those scores indicative of their actual abilities), would generally be greater than they would for sighted individuals. The graduate program's decision to admit or place a visually impaired student would, therefore, be subject to a greater likelihood of error than for nonhandicapped applicants.

The two indices most often used to assess test reliability are the reliability coefficient and the standard error of measurement (SEM). By definition, the reliability coefficient is affected by the amount of test score dispersion in a group, with smaller variances tending to produce smaller reliability coefficients. Because of this sensitivity to within-group homogeneity, the reliability coefficient is limited as a comparative measure of precision

across groups. (It retains utility, however, as an index of the test's ability to separate individuals within a given group.) The standard error of measurement is relatively unaffected by score variance. Hence, it is better suited to the comparison of measurement accuracy across populations.

For an objective test, estimates of reliability should typically take account of several sources of variation including that due to differences in the samples of items used to assess scholastic ability and differences in the occasions on which ability is assessed. To account for error due to the latter source, two administrations on different days need to be arranged for the same examinee sample. Because arranging two administrations is often impractical, reliability indices that can be derived from only one administration more commonly are used. These indices, the most popular of which is coefficient alpha, have been found to closely approximate those derived from repeated testings (Swineford, 1973).

Table 9 presents alpha reliability coefficients and standard errors of measurement for handicapped and nonhandicapped examinees taking C-3DGR3. For the nonhandicapped students, reliability coefficients for the Verbal, Quantitative, and Analytical scales are .90, .91, and .85, respectively. Coefficients for the disability groups fall consistently within a few points of these values, with the exception of one case in which the coefficient is somewhat higher: for the visually impaired-

large type (VIL) sample the Analytical scale coefficient is .91. This higher value is likely due to the unusually large variation in Analytical scores found in this group (see Table 4).

Insert Table 9 about here

Standard errors of measurement are presented in raw score units. For nonhandicapped examinees taking the 76-item Verbal section, the 60-item Quantitative section, and the 47-item Analytical section, the raw-score SEMs are 3.59, 3.07, and 2.95, respectively. Without exception, the SEMs for the handicapped groups are virtually identical to these values, differing in the worst case by only nine hundredths of an item.

Because coefficient alpha may somewhat overestimate reliability when a test is speeded (Anastasi, 1976), Table 10 presents reliability indices and standard errors of measurement calculated in a different manner. These calculations are based on the fact that sections within General Test scales are constructed according to similar specifications: each section contains the same item types, the same number and difficulty of items within a type, and the same ordering of types. In essence, sections can be viewed as parallel forms of the same test. By correlating scores on each of these parallel forms, and correcting (via the Spearman-Brown formula) for the fact that each parallel

form is half the length of the full scale, a reasonably accurate estimate of reliability can be obtained. Like coefficient alpha, however, the primary source of error assessed will be variation in the sample of items used. Variation due to differences in the occasions on which ability is assessed will not be treated as an error source since parallel forms were administered during the same sitting.

Insert Table 10 about here

Review of Table 10 shows reliability coefficients and standard errors that suggest only slightly less precision than those based on coefficient alpha. With respect to the reference group, the reliability coefficients for the handicapped groups again fall within a few points on all scales. The standard errors of measurement for the disability groups, which are somewhat more variable than those based on coefficient alpha, also generally are comparable to the reference indices. The most discrepant index (in the direction of indicating more error) is for the visually impaired large-type (VIL) group on Analytical. This index is four tenths of an item greater than the reference value.

Unexpected Differential Performance

The concept of unexpected differential performance is derived from the notion that items on a unidimensional test

should measure the same construct for different groups of examinees (Shepard, 1982). Items found to measure different constructs across groups are considered to be biased in the sense that, for some groups, they may be assessing factors irrelevant to the purpose of assessment. If found in any number, such items may unfairly lower (or raise) the test scores of a group. In many cases, however, biased items are found to, in the aggregate, equally affect different groups, cancelling any overall advantage or disadvantage that might otherwise be afforded (Berk, 1982; Shepard, 1982). Still, the identification of such items is important, for it alerts test developers to the kinds of questions that should be removed from future test revisions, or at least not disproportionately added lest the balance of questions favoring and disfavoring groups be destroyed.

Most methods of detecting items that operate differently across groups consider an item to be biased if groups of equal ability perform differently on it. This definition of item bias makes most sense if it can be assumed that the test or subtest under investigation is basically unidimensional (Shepard, 1982). If the measure can be considered safely to be unidimensional, then differences in performance on an item that remain after standing on the dimension has been accounted for must be due to irrelevant sources.

A second common characteristic of item bias methods is that total test score is used as a proxy for ability level

(Shepard, 1982). If all items in the test measure the same irrelevant construct for one group, it is possible that no indication of bias will appear; no item will stand out because it measures something different from the others. Item bias methods cannot, therefore, detect pervasive bias in a test because they lack an external criterion. As such, the study of item bias can be only one part of a comprehensive investigation of a test's fairness. Such an investigation should also consider the test's factor structure, to see if the test actually measures the same general construct across groups, and its relationship with external variables, to ensure that relevant criteria are predicted with equal accuracy.

To detect the possible presence of item bias, investigators have typically contrasted the performance of each group on each individual item with that of some reference group (e.g., see Kulick, 1984). In studies in which several groups are involved, this method has practical disadvantages. First, it necessitates attention to a large number of item performances. In the present study, 186 items and three handicapped groups would generate 558 contrasts. Second, even in groups in which bias is known not to exist (e.g., two random samples drawn from the same population), statistical techniques will identify by chance alone some small proportion of items as biased (Sinnott, 1980). Though proportionately small, the number of items so identified can, in absolute terms, be large: given a .05

significance criterion and 558 contrasts, some 28 items would be expected to show evidence of differential operation by chance alone. These items would, of course, be mixed with other correctly identified items. Separating the two item groups through content analysis would take substantial time and effort, and in some cases be unsuccessful because the underlying causes for differential operation are frequently unclear (Scheuneman, 1982).

To avoid some of these problems, a two-stage method was used to detect the possible presence of broad General Test item types that operate differently across groups. First, items were organized into logical clusters. Cluster structures were based both on those characteristics that might prove unusually troublesome for particular groups of handicapped examinees and on groupings typically used in the development of the General Test. Next, the performance of these clusters was investigated. Items belonging to deviantly-operating clusters were then individually studied to determine if the cluster itself defined a potentially biased item type or, alternatively, if only a few aberrant items accounted for the unusual cluster performance.

Item clusters. The rationale behind the study of item clusters is generally the same as that used for items: on a test measuring a single construct, a cluster should be of equal difficulty for different groups of the same ability. If not, the cluster is measuring something in addition to the intended construct in the groups.

To examine the performance of clusters, Verbal section items were divided by type into the four formats used in the test: antonyms, analogies, sentence completion, and reading comprehension. Item type was selected as the basis for forming clusters because previous studies have suggested that particular Verbal item types may be differentially difficult for handicapped as well as other special populations. For example, visually impaired students taking the GRE are reported to perceive reading comprehension items as posing particular difficulties because of their handicap (Ragosta & Kaplan, in press).

Once item-type clusters were formed, the regression of each cluster score on the total Verbal score for nonhandicapped students taking C-3DGR3 (i.e., the standardization group) was computed. This regression provided a prediction of performance for the standardization group on each item-type cluster for each Verbal score level. Using the Verbal mean for each disability group in turn, the predicted cluster score for a nonhandicapped group of the same total ability was generated. The predicted cluster mean for the nonhandicapped group was then subtracted from the handicapped group's actual cluster mean, yielding a positive residual if the disabled students did better than the reference group and a negative one when performance was lower than the reference group. Finally, this residual was divided by the cluster standard deviation for the disability group. A meaningful departure from the expected difficulty

of the cluster was said to exist for a group if the standardized residual exceeded an absolute value of .2 standard deviations. This .2 standard deviation criterion has been previously suggested as a minimum for identifying the presence of meaningful effects in the social sciences (Cohen, 1969). In addition, in a previous study the criterion proved sensitive enough to identify clusters containing items showing only minimal differential effects (Bennett, Rock, & Kaplan, 1985).

Table 11 presents standardized residuals for each disability group's performance on the four Verbal item types. As can be seen, most values are below .1 standard deviations in magnitude and no value exceeds .2 standard deviations. The only value that comes closest to the .2 criterion is for the visually impaired-large type sample on the antonyms cluster, which is somewhat differentially easy for this group.

Insert Table 11 about here

For the Quantitative section, standardized residuals for the item clusters were calculated in a way similar to that for Verbal with two exceptions. First, the cluster scores of nonhandicapped examinees were regressed on total Quantitative (instead of total Verbal) score to obtain a prediction of expected cluster performance for nonhandicapped groups with the same total scores as the disability

groups. Second, several overlapping cluster structures were studied based on item type, content, the presence of graphical material, and reading load. More than one structure was studied because Quantitative items appear, at least on the surface, to require a broader constellation of basic skills for solution, thereby allowing more room for bias. For example, in addition to reasoning ability, some Quantitative items require the visual skills needed to read graphs, tables, or special symbols, or to manipulate figures in space; for visually impaired examinees, these items may be more a measure of visual-spatial than math reasoning abilities. Other items, such as word problems, require reading skill. Because visually impaired students report particular difficulty with items involving reading (Ragosta & Kaplan, in press), word problems may not accurately assess math reasoning ability in this group.

For the first analysis, three clusters based on item type were employed: quantitative comparisons, discrete quantitative, and data interpretation. Table 12 presents standardized residuals for these cluster structures. As review of the table suggests, no residual exceeds the .2 criterion and only one is larger than .1.

Insert Table 12 about here

The second cluster structure investigated involved three item groups based on test content. Items were

assigned to clusters according to three of the four test development content classifications: arithmetic, algebra, and geometry. Items belonging to the fourth classification, data interpretation, were omitted as these were tested in the previous item-type structure.

Standardized residuals for the three content clusters are presented in Table 13. As inspection of the table bears out, the residuals for this cluster structure are similar to those for the previous one: most are below .1 and none exceeds the .2 criterion.

Insert Table 13 about here

The third structure analyzed included four clusters based on the presence or absence of graphics. To form clusters, items were divided into graphics (i.e., items including tables or figures) and nongraphics, or text, questions. Items were considered to involve graphics only if a graphic was actually presented. To limit the overlap between cluster and total scores, the graphics and text groupings were further divided according to section. Graphics items appearing in the first Quantitative section of the test were assigned to one cluster and those appearing in the second section allocated to another. This additional split produced two graphics and two text clusters.

Standardized residuals for these clusters are presented in Table 14. As the table shows, no residual exceeds .1.

The absence of differential effects for graphics is encouraging. It suggests that items containing figural material do not generally interfere with the accurate measurement of quantitative ability for those visually impaired examinees who take the General Test in regular print or large type formats.

Insert Table 14 about here

The final Quantitative cluster structure examined involved three groupings based on reading load: nonreading, minimal reading, and reading. An item was placed in the reading category if it contained more than one line of text. Minimal reading items were those with approximately one line of text or less, while nonreading items contained no words, only mathematical symbols and numerals. Written directions introducing the various item types at the beginning and middle of each of the two Quantitative sections were not included in the analysis as the amount of reading entailed was constant for all items.

As can be seen from Table 15, standardized residuals for the reading load clusters are only slightly larger than those found for previous structures. Four residuals exceed .1 and one almost reaches the .2 criterion. This latter residual, for visually impaired-large type students taking the nonreading cluster, suggests slight differential difficulty for the group.

Insert Table 15 about here

For the Analytical scale, standardized residuals were calculated as for the other scales, except that cluster score was regressed on total Analytical score. On this scale, a single cluster structure based on item type was analyzed. Items were first divided into analytical and logical reasoning types. These groupings were then subdivided according to General Test section, resulting in the formation of two analytical and two logical reasoning clusters. As Table 16 shows, all residuals are well below the .2 criterion with the singular exception of the logical reasoning I cluster, which is differentially easy for visually impaired-large type (VIL) students.

Insert Table 16 about here

In sum, the analysis of item clusters has identified only a single effect of a magnitude large enough to warrant closer study. This effect was associated with the Analytical section of the General Test and was positive, denoting that the cluster were unexpectedly easy for the handicapped group.

Individual items. The identification of item clusters can be considered the first, or screening, stage in a two-tiered procedure for detecting broad item classes that

appear to operate differently for handicapped and nondisabled populations. Therefore, after screening the Verbal, Quantitative, and Analytical item clusters and identifying one cluster that appeared to operate differently, a second, more focused methodology was applied. This methodology was designed to detect individual items that seemed to contribute significantly to the finding of differential difficulty for the cluster. In so doing, the methodology indicated the extent to which cluster effects were due to a few isolated items or, alternatively, to the preponderance of items composing the type. In addition, the methodology was meant to provide rigorous statistical tests of the implicit assumption that the relationship between total score and the probability of passing a given item is the same in the standard and handicapped populations.

To accomplish these goals, logistic regression was used to analyze performance on those items composing the identified cluster. Within the identified cluster, the item performance of the handicapped group was contrasted with the standardization population (i.e., nonhandicapped students taking C-3DGR3) to determine if the expectations of passing a given item were equivalent across groups. Item performances were compared after controlling for both total score and sex group membership. Sex group membership was considered because it has been associated with differential performance on verbal and mathematics items and because the distribution of sexes across groups varied somewhat in the

present study. Hence, differences in sex distribution might account partially for the observed cluster performance. Logistic regression also was used to compare the equality of the slopes of an item performance on total test score for handicapped and nondisabled groups. This latter comparison indicated the extent to which an item evidenced differential operation as a function of ability level (e.g., no differential operation for low-scoring handicapped examinees but differential difficulty for high-scoring ones).

More formally, in the standardization population, c, the probability, P, of passing item i given a total score X = X' is:

$$P_{ci} = P(x_i = 1 \mid X = X')$$

where x is a 0 or 1 score obtained on item i. Similarly, in any given handicapped group, h,:

$$P_{hi} = P(x_i = 1 \mid X = X')$$

The question to be answered is whether:

$$P_{ci} - P_{hi} \neq 0?$$

The logistic regression model first estimates the unknown regression parameters (i.e., the B's) in the following equation:

$$\log (P / (1 - P)) = B_0 + B_1 D + B_2 S + B_3 X$$

where P is the probability of passing a given item, D and S are dummy variables indicating handicapped vs. nonhandicapped and sex group membership, respectively, and X is the total test score.

To simplify the analysis, the test score distribution of the handicapped group was partitioned into thirds. The scaled scores bounding each third were 200-450, 460-620, and 630-800. These boundaries were then used to establish score groups for the reference sample, thereby "matching" the handicapped and reference groups on total score.

Table 17 presents results for the visually impaired-large type and standardization groups on items belonging to the logical reasoning I cluster. For each item, the table indicates whether each of three variables--presence of handicap, sex, and score level--is related to item performance when the other two variables are held constant. As the table indicates, score level is a significant--and expected--predictor of passing each of the four items when the presence of handicap and sex are controlled; the probability of passing an item should be dependent on developed ability, which is presumed to be represented by total Analytical score. For two of the four items, neither sex group nor membership in the visually impaired sample is related to performance when the other variables are held constant. Along with the fact that the logical reasoning II cluster showed no evidence of differential operation, this

result suggests that logical reasoning items per se present no particular advantage or disadvantage to visually impaired-large type examinees.

Insert Table 17 about here

The other two items in the cluster, however, do offer evidence of differential operation: when total score level is controlled, membership in the handicapped group (holding sex constant) and sex (controlling for the presence of handicap) both are related to item performance. Table 18 presents the observed proportions of examinees passing item 23 for each score level and sex group within the handicapped and nondisabled samples. Because of the extremely small sample sizes in the cells for the visually impaired group, the observed proportions, and the consequent differences in proportions between the handicapped and nondisabled group and within the handicapped group, may not be stable. It is, therefore, suggested that differences in proportions be at least .1 before a result is considered practically meaningful.

Insert Table 18 about here

As the table indicates, the item is differentially easy for visually impaired individuals: within total score groups and regardless of sex, greater proportions of these

examinees pass the item than their reference group counterparts. On average, the proportion of visually impaired examinees passing the item is 16 points greater for males and 21 points greater for females.

The sex effect shown by the item is most apparent within the lower and middle score groups of the reference sample. Across all score groups in this sample, the proportion of males passing the item is 10 points higher than for females. Within the handicapped sample, the sex effect appears marginal.

Table 19 presents comparable data for item 24. With the single exception of low-scoring males, this item too seems differentially easy for visually impaired examinees. On average, the proportions passing the item are 11 and 14 points higher than the reference group for males and females, respectively.

The effect of sex on passing the item is more complex. In the reference group, males are favored marginally. In the handicapped group, an interaction appears to exist, with low scoring females finding the item differentially easy and high scoring females finding it unexpectedly hard. The extremely small sample sizes in these cells, however, suggest that this finding may not be dependable.

Insert Table 19 about here

Why do these items appear differentially easy for visually impaired students? The text of each item is given in Figure 4. From review of the items' content no indication of why the questions should favor visually impaired examinees is immediately apparent. These items are, however, the last items in the first Analytical section of the General Test. Since 95% of visually impaired individuals complete the section, compared with 83% of reference examinees (see Table 8), the finding may be due to extended time.

Insert Figure 4 about here

Why is item 23 differentially difficult for females? This item requires reasoning with scientific content. Males often are more interested in science at earlier ages than females and are more likely to have been physical science majors as undergraduates (Smith, 1985). As such, males may have developed greater facility and may be more comfortable with such material than females of equal developed ability.

In summary, the items composing the logical reasoning I cluster were individually studied to determine whether this item type represented a broad class that operated differently for visually impaired examinees taking the large-type edition of the General Test. Of the four items composing the cluster, only two showed statistically significant evidence of differential operation. Coupled

with the fact that a second logical reasoning cluster showed no indication of differential operation, the finding of cluster deviancy for logical reasoning would appear to be attributable to these two isolated items and not to the general functioning of the item type. Although the reasons for the differential easiness of these two isolated items could not be definitively determined, the placement of these items at the end of the test section coupled with the extra time given visually impaired examinees, may account for the obtained result.

Summary

This study has investigated the psychometric characteristics of the GRE General Test for three handicapped groups. Data on these characteristics are central to evaluating the accuracy of scores for measuring the developed scholastic abilities of disabled examinees.

With respect to performance level, the visually handicapped groups achieved mean scores that approximated those of nondisabled students. On the three GRE scales, examinees taking the large-type administration scored between -.11 and .28 standard deviations from the corresponding means for all students taking the test during the same period, while visually handicapped students who took the standard, national administration performed from .13 to .10 standard deviations below the comparison group. Physically handicapped-national administration examinees

scored lower, from .43 to .03 units below nonhandicapped comparison students.

The second psychometric characteristic examined related to whether disabled students finish the test to the same degree as their nonhandicapped peers. With respect to the reference sample, physically handicapped and visually impaired students taking timed, national administrations were slightly less likely to complete selected test sections. For both groups, these differences were associated with the Analytical sections, though differences were also found for the former group on one of the Quantitative sections. It is uncertain whether these findings indicate that the developed analytical and quantitative abilities of handicapped students are being underestimated by timed administrations. Further research--in particular, predictive and construct validity investigations--should provide a clearer indication of the accuracy of scores from these administrations.

Score precision was assessed using two methods of calculating the reliability coefficient and the standard error of measurement. The reliability of the General Test was found to be comparable to the reference population for all handicapped groups.

The final psychometric characteristic studied was unexpected differential performance. This characteristic was studied to identify broad item classes that may not measure the ability assessed by the overall test.

Differential performance was evaluated through a two-stage procedure in which the operation of groups of items was investigated, followed by a study of the individual items composing deviantly-operating groups. One item group and two individual items within this group were identified as potentially problematic for visually impaired students taking the large-type administration. This differential item behavior was hypothesized to result from the placement of the items at the end of the test section coupled with the extra time afforded visually handicapped students. Further research might explore the question of whether these examinees are advantaged unduly by the provision of extended time.

In conclusion, the reliability and item behavior of General Test form C-3DGR3 appear comparable for the disabled and nonhandicapped groups taking part in this investigation. Though somewhat less comparable, test completion and test performance levels (which are closely linked) also were similar across groups. That the psychometric characteristics of the test are similar across populations lends support to the notion that General Test scores are accurate and fair indicators of the developed scholastic abilities of visually and physically disabled students. Further evidence from factor analyses and predictive validity studies should add knowledge about the meaning of these scores for handicapped examinees.

References

Anastasi, A. (1976). Psychological testing (fourth edition). New York: MacMillan.

Bejar, I. I. (1985). Test speededness under number-rights scoring: An analysis of the Test of English as a Foreign Language. Princeton, NJ: Educational Testing Service.

Bennett, R. E., & Ragosta, M. (1984). A research context for studying admissions tests and handicapped populations. Princeton, NJ: Educational Testing Service.

Bennett, R. E., Ragosta, M., & Stricker, L. J. (1984). The test performance of handicapped people (RR 84-32). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1985). The psychometric characteristics of the SAT for nine handicapped groups (RR-85-49). Princeton, NJ: Educational Testing Service.

Berk, R. A. (1982). Introduction. In R. A. Berk (Ed), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Burton, N. W., & Turner, N. J. (1983). Effectiveness of the Graduate Record Examinations for predicting first-year grades: 1981-82 summary report of the Graduate Record Examinations Validity Study Service. Princeton, NJ: Educational Testing Service.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Conrad, L., Trismen, D., & Miller, R. (Eds). (1977). Graduate Record Examinations technical manual. Princeton, NJ; Educational Testing Service.

Educational Testing Service. (1984). GRE 1984-85 Information Bulletin. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1980). Test use and validity. Princeton, NJ: Author.

Goodison, M. B. (1983). A summary of data collected from Graduate Record Examinations test takers during 1981-1982 (Data Summary Report #7). Princeton, NJ: Educational Testing Service.

Kulick, E. (1984). Assessing unexpected differential item performance of black candidates on SAT Form CSA6 and TSWE Form E33 (SR-84-80). Princeton, NJ: Educational Testing Service.

Ragosta, M., & Kaplan, B. A. (In press). A survey of handicapped students taking special test administrations of the SAT and GRE. Princeton, NJ: Educational Testing Service.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Sherman, S. W., & Robinson, N. M. (Eds). (1982). Ability testing of handicapped people: Dilemma for government, science, and the public. Washington, D. C.: National Academy Press.

Sinnott, L. T. (1980). Differences in item performance across groups (RR 80-19). Princeton, NJ: Educational Testing Service.

Smith, H. R. (1984). A summary of data collected from Graduate Record Examinations test takers during 1982-1983 (Data Summary Report #8). Princeton, NJ: Educational Testing Service.

Smith, H. R. (1985). A summary of data collected from Graduate Record Examinations test takers during 1983-1984 (Data Summary Report #9). Princeton, NJ: Educational Testing Service.

Swineford, F. (1973, February). An assessment of the Kuder-Richardson Formula (20) reliability estimate for moderately speeded tests. Paper presented at the meeting of the National Council for Measurement in Education, New Orleans, Louisiana.

Footnote

1. Relative to the general population, students taking the GRE General Test constitute a highly selected and unusually able group. Handicapped students taking the GRE also are likely to be unusually able. As such, this group should not be regarded as being representative of disabled people in general nor of disabled college students.

Table 1

Numbers of Students Taking Administrations of GRE General Test Form C-3DGR3[a]

Type of Administration	Handicapped Group[b]						
	VI	HI	PH[c]	LD	MH	O	T
ALTERNATE FORMAT							
Braille	39	0	0	0	0	0	39
Large type	151	0	8	3	13	4	185
Cassette	1	0	0	0	0	0	1
Multiple	59	0	0	16	1	0	76
REGULAR TYPE W/ EXTRA TIME	6	1	10	8	3	1	29
STANDARD NATIONAL	188	71	108	35	10	125	537
UNKNOWN	4	0	0	2	0	0	6
TOTAL	454	72	126	64	27	130	873

[a] Students for whom English may be a second language are excluded.

[b] VI = visually impaired; HI = hearing impaired; PH = physically handicapped; LD = learning disabled; MH = multiply handicapped; O = other; T = total.

[c] Includes neurologically or other orthopedically impaired, paraplegic, and permanent physical handicap classifications.

Table 2

Disability Groups Selected for Study

Acronym	Group	N
PHN	Physically handicapped students taking regular type test in a timed, national administration	108
VIL	Visually impaired students taking large-type test in a special, extended-time administration	151
VIN	Visually impaired students taking regular type test in a timed, national administration	188

Table 3

Performance of Reference Students Taking National Administrations of the GRE General Test[a]

Scale	Group[b] NHF (N=20,499)	NHA (N=441,654)
Verbal		
Mean	497	499
SD	115	117
Quantitative		
Mean	524	526
SD	134	133
Analytical		
Mean	520	524
SD	126	125

[a] Students for whom English may be a second language are excluded.

[b] NHF = nonhandicapped examinees taking test form C-3DGR3 in October 1981 and April 1982; NHA = nonhandicapped examinees taking all General Test forms offered from October 1981 through June 1984.

Table 4

General Test Performance

of Handicapped Examinees

Group	Mean	Standard Deviation	Difference in SD Units From Reference Group [a]
Verbal Scale			
PHN	493	117	-.03
VIL	513	109	.12
VIN	486	120	-.10
Quantitative Scale			
PHN	467	130	-.43
VIL	512	137	-.11
VIN	507	131	-.13
Analytical Scale			
PHN	481	120	-.31
VIL	559	154	.28
VIN	503	135	-.13

[a] Reference group for VIL is all nonhandicapped students taking the General Test from October 1981 through June 1984. Reference group for VIN and PHN consists of nonhandicapped students taking C-3DGR3 in contemporaneous national administrations. Differences are expressed in SD units of the appropriate reference group.

Table 5

Group Composition by Sex of

General Test Examinees

Group[a]	Percent Male	Percent Female
PHN	51	47
VIN	43	56
NHF	42	57
VIL	35	64
NHA	44	56

[a] NHF = nonhandicapped students taking form C-3DGR3 in October 1981 and April 1982; NHA = nonhandicapped students taking all forms of the General Test from October 1981 through June 1984.

Table 6

Percent of General Test Examinees

Reporting Different Undergraduate Majors

Major[b]	Group[a] PHN	VIN	NHF	VIL	NHA
Humanities	19	15	14	18	15
Social Sciences	50	45	43	57	41
Biological Sciences	12	21	21	9	21
Physical Sciences	16	11	16	7	17
Other/Unknown	5	9	6	10	6

[a] NHF = nonhandicapped students taking GRE General Test form C-3DGR3 in October 1981 and April 1982; NHA = nonhandicapped students taking all forms of the General Test from October 1981 through June 1984.

[b] Classifications based on background information questions included on the GRE registration form.

Table 7

Extent of General Test Completion for Handicapped Examinees[a]

Group	Verbal Section I: Ratio of % Completing Test	Verbal Section I: Ratio of % Completing 75% Test	Verbal Section II: Ratio of % Completing Test	Verbal Section II: Ratio of % Completing 75% Test
PHN	1.04	1.01	1.00	.99
VIL	1.08	1.02	1.04	1.01
VIN	.98	1.00	1.01	.99
	Quantitative Section I		**Quantitative Section II**	
PHN	.89	.95	.90	1.00
VIL	1.08	1.01	1.00	1.01
VIN	.93	1.00	1.02	1.00
	Analytical Section I		**Analytical Section II**	
PHN	.89	.91	.93	.94
VIL	1.14	1.04	1.10	1.01
VIN	.89	.96	.89	.95

[a] Tabled values are the ratio of handicapped students to nonhandicapped students completing the section or section part.

Table 8

Percent of Examinees Completing the General Test

Group	Verbal Section I	Verbal Section II
PHN	92	94
VIL	95	97
VIN	86	94
NHF	89	93

	Quantitative Section I	Quantitative Section II
PHN	78	82
VIL	94	91
VIN	81	93
NHF	87	92

	Analytical Section I	Analytical Section II
PHN	74	78
VIL	95	91
VIN	74	74
NHF	83	83

Table 9

Internal Consistency Reliability of

General Test Form C-3DGR3

Group[a]	Coefficient Alpha	Standard Error of Measurement
Verbal Scale (76 items)		
PHN	.91	3.61
VIL	.89	3.60
VIN	.91	3.62
NHF	.90	3.59
Quantitative Scale (60 items)		
PHN	.89	3.16
VIL	.90	3.10
VIN	.90	3.12
NHF	.91	3.07
Analytical Scale (47 items)		
PHN	.83	2.97
VIL	.91	2.89
VIN	.86	2.95
NHF	.85	2.95

[a] NHF = nonhandicapped students taking form C-3DGR3.

Table 10

Parallel Forms Reliability

of General Test Form C-3DGR3

Group[a]	Reliability Coefficient	Standard Error of Measurement
Verbal Scale (70 items)		
PHN	.88	4.07
VIL	.88	3.70
VIN	.90	3.80
NHF	.89	3.85
Quantitative Scale (60 items)		
PHN	.89	3.17
VIL	.90	3.16
VIN	.88	3.37
NHF	.90	3.18
Analytical Scale (47 items)		
PHN	.81	3.12
VIL	.85	3.64
VIN	.83	3.28
NHF	.82	3.20

[a] NHF = nonhandicapped students taking form C-3DGR3.

Table 11

Unexpected Differential Performance in SD Units on Verbal Item-Type Clusters[a]

Group	Antonyms (22 items)	Analogies (18 items)
PHN	.10	.06
VIL	.17	-.02
VIN	-.01	.04

Group	Reading Comprehension (22 items)	Sentence Completion (14 items)
PHN	-.10	-.04
VIL	-.07	-.07
VIN	-.02	-.01

[a] Values are computed by subtracting the predicted mean cluster score for a handicapped group from its actual mean cluster score and dividing the result by the cluster standard deviation for the group. Negative values indicate performance that is lower than expected.

Table 12

Unexpected Differential Performance in SD Units on Quantitative Item-Type Clusters[a]

Group	Quantitative Comparisons (30 items)	Discrete Quantitative (19 items)
PHN	.00	-.01
VIL	-.08	.04
VIN	-.02	.02

Group	Data Interpretation (10 items)
PHN	.04
VIL	.14
VIN	.00

[a] Values are computed by subtracting the predicted mean cluster score for a handicapped group from its actual mean cluster score and dividing the result by the cluster standard deviation for the group. Negative values indicate performance that is lower than expected.

Table 13

Unexpected Differential Performance in SD Units on Quantitative Content Clusters[a]

Group	Arithmetic (17 items)	Algebra (16 items)
PHN	.07	-.08
VIL	.00	-.14
VIN	.00	.03

Group	Geometry (16 items)
PHN	.00
VIL	.04
VIN	-.01

[a] Values are computed by subtracting the predicted mean cluster score for a handicapped group from its actual mean cluster score and dividing the result by the cluster standard deviation for the group. Negative values indicate performance that is lower than expected.

Table 14

Unexpected Differential Performance

in SD Units on Quantitative Graphics-Load Clusters[a]

Group	Graphics I (10 items)	Graphics II (11 items)
PHN	.01	.00
VIL	.07	.01
VIN	-.04	.02

Group	Text I (20 items)	Text II (18 items)
PHN	.05	-.04
VIL	.02	-.07
VIN	.01	.01

[a] Values are computed by subtracting the predicted mean cluster score for a handicapped group from its actual mean cluster score and dividing the result by the cluster standard deviation for the group. Negative values indicate performance that is lower than expected.

Table 15

Unexpected Differential Performance

in SD Units on Quantitative Reading-Load Clusters [a]

Group	Reading (31 items)	Minimal Reading (10 items)
PHN	.10	-.04
VIL	.13	.02
VIN	.02	-.04

Group	Non-Reading (18 items)
PHN	-.10
VIL	-.19
VIN	.01

[a] Values are computed by subtracting the predicted mean cluster score for a handicapped group from its actual mean cluster score and dividing the result by the cluster standard deviation for the group. Negative values indicate performance that is lower than expected.

Table 16

Unexpected Differential Performance

in SD Units on Analytical Item-Type Clusters [a]

Group	Analytical Reasoning I (18 items)	Analytical Reasoning II (19 items)
PHN	-.06	.02
VIL	.07	-.12
VIN	-.04	.05

Group	Logical Reasoning I (4 items)	Logical Reasoning II (6 items)
PHN	.03	.07
VIL	.23	.03
VIN	.05	-.07

[a] Values are computed by subtracting the predicted mean cluster score for a handicapped group from its actual mean cluster score and dividing the result by the cluster standard deviation for the group. Negative values indicate performance that is lower than expected.

Table 17

t-Values for Three Variables Included in an Equation to Predict the Item Performance of Visually Impaired-Large Type and Nonhandicapped Examinees[a]

Logical Reasoning Item	Predictor Variable: Presence of Handicap	Sex	Score Level
I6	.53	-1.02	34.64***
I8	-1.08	-1.23	46.09***
I23	4.88***	-14.97***	45.80***
I24	3.64***	-3.27**	37.20***

** $p < .01$
*** $p < .001$

[a] t-values indicate the significance of each variable in predicting item performance when the other two variables are held constant.

Table 18

Proportions of Examinees Passing[a] Logical Reasoning I Item 23

Group[b]	Score Level: Low	Medium	High	Average
VIL				
Male	.59 (17)	.82 (17)	.94 (18)	.79
Female	.57 (30)	.77 (30)	.88 (33)	.74
NHF				
Male	.42 (2556)	.63 (4162)	.83 (1911)	.63
Female	.28 (4256)	.53 (5397)	.78 (2022)	.53
Average	.46	.69	.86	.67

[a] The total number of examinees responding to the item is presented in parentheses.

[b] NHF = nonhandicapped students taking form C-3DGR3.

Table 19

Proportions of Examinees Passing
Logical Reasoning I Item 24[a]

Group[b]	Score Level			
	Low	Medium	High	Average
VIL				
Male	.12 (17)	.53 (17)	.78 (18)	.47
Female	.27 (30)	.57 (30)	.61 (33)	.48
NHF				
Male	.19 (2556)	.32 (4162)	.56 (1911)	.36
Female	.18 (4256)	.29 (5397)	.54 (2022)	.34
Average	.19	.43	.62	.41

[a] The total number of examinees responding to the item is presented in parentheses.

[b] NHF = nonhandicapped students taking form C-3DGR3.

Figure 1

GRE General Test

Verbal Item Types

Analogies

Directions: In each of the following questions, a related pair of words or phrases is followed by five lettered pairs of words or phrases. Select the lettered pair that best expresses a relationship similar to that expressed in the original pair.

1. COLOR : SPECTRUM :: (A) tone : scale
(B) sound : waves (C) verse : poem
(D) dimension : space (E) cell : organism

Antonyms

Directions: Each question below consists of a word printed in capital letters followed by five lettered words or phrases. Choose the lettered word or phrase that is most nearly *opposite* in meaning to the word in capital letters. Since some of the questions require you to distinguish fine shades of meaning, be sure to consider all the choices before deciding which one is best.

5. DIFFUSE : (A) concentrate (B) contend
(C) imply (D) pretend (E) rebel

Sentence Completion

Directions: Each sentence below has one or two blanks, each blank indicating that something has been omitted. Beneath the sentence are five lettered words or sets of words. Choose the word or set of words for each blank that *best* fits the meaning of the sentence as a whole.

9. Early ——— of hearing loss is ——— by the fact that the other senses are able to compensate for moderate amounts of loss, so that people frequently do not know that their hearing is imperfect.

(A) discovery . . indicated
(B) development . . prevented
(C) detection . . complicated
(D) treatment . . facilitated
(E) incidence . . corrected

Figure 1 (Cont'd)

GRE General Test

Verbal Item Types

Reading Comprehension

Directions: The passage is followed by questions based on its content. After reading the passage, choose the best answer to each question. Answer all questions following the passage on the basis of what is *stated* or *implied* in the passage.

In the years following the Civil War, economic exploitation for the first time was provided with adequate resources and a competent technique, and busy prospectors were daily uncovering new sources of wealth. The coal and oil of Pennsylvania and Ohio, the copper and iron ore of Upper Michigan, the gold and silver, and the lumber and fisheries of the Pacific Coast provided limitless raw materials for the rising industrialism. The Bessemer process quickly turned an age of iron into an age of steel and created the great mills of Pittsburgh from which issued the rails for expanding railways. The reaper and binder, the sulky plow, and the threshing machine created a large scale agriculture on the fertile prairies. Wild grasslands provided grazing for immense herds of cattle and sheep; the development of the corn belt enormously increased the supply of hogs; and with railways at hand the Middle Border poured into Omaha and Kansas City and Chicago an endless stream of produce.

As the line of the frontier pushed westward, new towns were built, thousands of claims to homesteads were filed, and speculator and promoter hovered over the prairies like buzzards seeking their carrion. With rising land values money was to be made out of unearned increment, and the creation of booms was a profitable industry. The times were stirring, and it was a shiftless fellow who did not make his pile. If he had been too late to file on desirable acres, he had only to find a careless homesteader who had failed in some legal technicality and "jump his claim." Good bottom land could be had even by late-comers if they were sharp at the game.

This bustling America of 1870 accounted itself a democratic world. A free people had put away all aristocratic privileges and, conscious of power, had gone forth to possess the last frontier. But America's essential social philosophy, which it found adequate to its needs, was summed up in three words—preemption, exploitation, progress. Its immediate and pressing business was to dispossess the government of its rich holdings. Lands in the possession of the government were so much idle waste, untaxed and profitless; in private hands they would be developed. They would provide work, pay taxes, support schools, enrich the community. Preemption meant exploitation and exploitation meant progress.

It was a simple philosophy and it suited the simple individualism of the times. The Gilded Age knew nothing of enlightenment; it recognized only the acquisitive instinct. That much at least the frontier had taught the great American democracy; and in applying to the resources of a continent the lesson it had been so well taught, the Gilded Age wrote a profoundly characteristic chapter of American history.

12. According to the passage, increased corn production was mainly responsible for an increase in the

(A) number of sheep (B) output of farm implements
(C) supply of hogs (D) amount of pasture land
(E) number of cattle

Source: ETS. (1984). GRE 1984-85 Information Bulletin. Princeton, NJ: Author. Used by permission.

Figure 2

GRE General Test

Quantitative Item Types

Quantitative Comparison

Directions: Each of the following questions consists of two quantities, one in Column A and one in Column B. You are to compare the two quantities and choose

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be determined from the information given.

Note: Since there are only four choices, NEVER MARK (E).

Common Information: In a question, information concerning one or both of the quantities to be compared is centered above the two columns. A symbol that appears in both columns represents the same thing in Column A as it does in Column B.

	Column A	Column B
15.	9.8	$\sqrt{100}$

Data Interpretation

Questions 28-30 refer to the following table:

PERCENT CHANGE IN DOLLAR AMOUNT OF SALES IN CERTAIN RETAIL STORES FROM 1977 TO 1979

Store	Percent Change From 1977 to 1978	Percent Change From 1978 to 1979
P	+10	−10
Q	−20	+9
R	+5	+12
S	−7	−15
T	+17	−8

28. In 1979 which of the stores had greater sales than any of the others shown?

(A) P (B) Q (C) R (D) S
(E) It cannot be determined from the information given.

Discrete Quantitative

Directions: Each of the following questions has five answer choices. For each of these questions, select the best of the answer choices given.

22. The average of x and y is 20. If z = 5, what is the average of x, y, and z?

(A) $8\frac{1}{3}$ (B) 10 (C) $12\frac{1}{2}$ (D) 15 (E) $17\frac{1}{2}$

Source: ETS. (1984). GRE 1984-85 Information Bulletin. Princeton, NJ: Author. Used by permission.

Figure 3

GRE General Test

Analytical Item Types

Logical Reasoning

47. If Ramón was born in New York State, then he is a citizen of the United States.

The statement above can be deduced logically from which of the following statements?

(A) Everyone born in New York State is a citizen of the United States.
(B) Every citizen of the United States is a resident either of one of the states or of one of the territories.
(C) Some people born in New York State are citizens of the United States.
(D) Ramón was born either in New York or in Florida.
(E) Ramón is a citizen either of the United States or of the Dominican Republic.

Analytical Reasoning

To apply to college a student must see the school counselor, obtain a transcript at the transcript office, and obtain a recommendation from Teacher A or Teacher B.
A student must see the counselor before obtaining a transcript.
The counselor is available only Friday mornings and Tuesday, Wednesday, and Thursday afternoons.
The transcript office is open only Tuesday and Wednesday mornings, Thursday afternoons, and Friday mornings.
Teacher A is available only Monday and Wednesday mornings.
Teacher B is available only Monday afternoons and Friday mornings.

40. A student has already seen the counselor and does not care from which teacher she obtains her recommendation. Which of the following is a complete and accurate list of those days when she could possibly complete the application process in one day?

(A) Friday (B) Monday, Wednesday
(C) Monday, Friday (D) Wednesday, Friday
(E) Monday, Wednesday, Friday

Source: ETS. (1984). GRE 1984-85 Information Bulletin. Princeton, NJ: Author. Used by permission.

Figure 4

Items Showing Evidence of Differential Operation

for Visually Impaired - Large Type Examinees

SECTION V

Time—30 minutes

24 Questions

Directions: Each question or group of questions is based on a passage or set of conditions. In answering some of the questions, it may be useful to draw a rough diagram. For each question, select the best answer choice given.

23. The newly discovered moon of Jupiter has an orbit around the planet that is well within the Roche limit for Jupiter.

The Roche limit is the distance within which satellites with a liquid core break up because of the strength of a planet's gravitational pull.

Which of the following conclusions can most properly be drawn from the information above?

(A) The newly discovered satellite has begun to break up.
(B) The object called a newly discovered moon of Jupiter is not in fact a moon of Jupiter.
(C) The newly discovered moon of Jupiter does not have a liquid core.
(D) The Roche limit for Jupiter is farther away from the planet than is the Roche limit for other planets.
(E) Jupiter has a weak gravitational pull.

24. Choose the most logical completion for the following paragraph.

A pluralist society takes into account the claims of recognized groups, with their conflicting interests, and attempts to mediate between them. But no group that is not sufficiently well organized to be recognized can make its claims count for anything at all. Its advocates will be treated as extremists or foreign agents. Thus the net effect of government action taken along pluralist principles will be to _____.

(A) strike the best possible compromises
(B) suppress the freedom of speech
(C) give everyone part of what he or she claims
(D) reinforce existing social structures
(E) resolve conflicts of interest

Source: Educational Testing Service. Copyright 1981. Used by permission.

Appendix

Annotated Bibliography of Previous Reports of the "Studies of Admissions Testing and Handicapped People"

The following previous reports from "Studies of Admissions Testing and Handicapped People" are available upon request from Educational Testing Service, Research Publications Unit-Room T143, Princeton NJ 08541:

#1 Bennett, R., and Ragosta, M. A Research Context for Studying Admissions Tests and Handicapped Populations, 1984. (ETS Research Report 84-31)

This is the first of a series of reports emanating from a four year research effort to further knowledge of admissions testing and handicapped people. The authors describe the legal and educational issues that gave rise to this research and the major questions to be addressed. They discuss the distinguishing characteristics of different types of disability and the complex definitional problems that hamper any simple method of classifying examinees by type of handicap.

#2 Bennett, R., Ragosta, M., and Stricker, L. The Test Performance of Handicapped People, 1984 (ETS Research Report 84-32)

The purpose of this report was to summarize existing research information concerning the performance of handicapped people on admissions and other similar tests. As a group, handicapped examinees scored lower than did the nonhandicapped. Among the four major groups examined, physically handicapped and visually impaired examinees were most similar to the nondisabled population. Hearing disabled students performed least well. Available studies of the SAT and ACT generally supported the validity of those tests for handicapped people, but it was confirmed that research to date has been quite limited and has not addressed many important questions.

#3 Bennett, R., Rock, D., and Kaplan, B. The Psychometric Characteristics of the SAT for Nine Handicapped Groups, 1985. (ETS Research Report 85-49)

In this study the main finding was that with the exception of performance level, the characteristics of the Scholastic Aptitude Test (SAT) were generally comparable for handicapped and nonhandicapped students. The analyses focused on level of test performance, test reliability, speededness, and extent of unexpected differential item performance on the SAT. Visually impaired students and those with physical handicaps achieved mean scores similar to those of students taking the SAT in national administrations, while learning disabled and hearing impaired students scored lower than their nondisabled peers. Analysis of individual items revealed only a few instances of differential item performance localized to visually imparied students taking the Braille test.

#4 Rock, D., Bennett, R., and Kaplan, B. The Internal Construct Validity of the SAT Across Handicapped and Nonhandicapped Populations, 1985. (ETS Research Report 85-50)

This study further investigated the comparability of SAT Verbal and Mathematical scores for handicapped and nonhandicapped populations. A two-factor model based on Verbal and Mathematical item parcels was posed and tested for invariance across populations. This model provided a reasonable fit in all groups, with the mathematical reasoning factor generally showing a better fit than the verbal factor. Compared with the nonhandicapped population, these factors tended to be less correlated in most of the handicapped groups. This greater specificity implies the increased likelihood of achievement growth in one area independent of the other and suggests that SAT Verbal and Mathematical scores be interpreted separately rather than as an SAT composite. Finally, there was evidence that the Mathematical scores for learning disabled students taking the cassette test may underestimate the reasoning ability of this group.

#5 Ragosta, M., and Kaplan, B. A Survey of Handicapped Students Taking Special Test Administrations of the SAT and GRE, 1986 (ETS Research Report, 86-5).

Disabled people were surveyed to obtain their views on the appropriateness of special test accommodations available for the Scholastic Aptitude Test (SAT) and the Graduate Record Examinations (GRE). More than nine out of ten respondents reported satisfaction with special test accommodations. A minority experienced dissatisfaction with the level of test difficulty or about specific shortcomings associated with test administrations. In comparing SAT and GRE administrations with accommodations normally provided in college testing, respondents reported that the admissions tests were more frequently offered in special versions and with extra time than were college tests.